Holonic Science: A New Framework for Benchmarking AI Scientists

Anonymous Author(s)
Affiliation
Address
email

Abstract

Science is a system defined in part by measurability. Claims made under its banner are trusted under the implicit understanding that they can be verified through measurement. Trustworthy science is therefore only possible when accurate and verifiable measurements of all aspects of a discovery or observation are possible. Recently, a new interloper has emerged in the form of AI scientists. Driven by companies such as Sakana AI and Google, these hybrid human-AI systems tasked with scientific discovery strive to augment and accelerate the current research paradigm by intelligently innovating upon and combining preexisting ideas. As researchers attempt to build collaborative workflows with AI scientists, the need for better measurements of their capabilities and limitations escalates. In this paper, we argue that the complexity of scientific research represents a significant challenge to AI scientist benchmarking attempts on account of construct validity issues. Scientific research tasks must be parseable by AI scientists, otherwise these in silico collaborators pose a significant epistemic risk to the trustworthiness of scientific research. To address this, we propose a new framework for designing benchmarks for AI scientists based on Arthur Koestler's concept of holons. Instead of benchmarking high-level human-interpretable tasks, we instead break them down and build specialized benchmarks at the LLM-executable level. These semantic sum of an AI scientist's performance on these benchmarks will then approximate performance on the original task. Our framework outlines key criteria for future benchmarks to avoid construct validity issues. We also exemplify the potential of our framework by prototyping a benchmark for attributional accuracy ultimately aimed at evaluating AI scientists on their ability to generate literature reviews.

1 Introduction

Science is a human-designed system that exists to engage with and explain the natural universe and thus is defined by our extent of measurement. As contributors to the system, the scientific community has evolved into a knowledge ecosystem joined by the mission that any claim can be verified due to its inherent measurability. Science is therefore only "scientific" when accurate and verifiable measurements are possible, for it is the measurability of scientific results that differentiates modern science from its medieval antecedents like alchemy and astrology. The critical subtlety here is that trustworthy science is only possible if all steps of the scientific process are verifiably measurable and documentable. However, this imperative now finds itself challenged by rising interest in utilizing hybrid human-AI systems known colloquially as 'AI scientists' to automate scientific discovery workflows.

AI systems have already proven to be useful inclusions in the pursuit of scientific discoveries. Both the 2024 Nobel Prize in Chemistry (1) and 2025 Nobel Prize in Physics (2) have proven the value

Submitted to the AI for Science Workshop (NeurIPS 2025). Do not distribute.

provided by AI models in performing complex computational scientific tasks. As such, it is no surprise that scientists have pursued even greater degrees of collaboration with AI systems, ultimately seeking to fully automate the whole scientific discovery workflow. Numerous works across multiple fields of study have sought to demonstrate that large language models (LLMs) are capable of performing scientific discovery tasks such as scientific contextualization (3), problem specification (4)(5)(6), hypothesis generation (7), experimental design (8)(9), and evaluation (10)(11). However, these attempts are still a far cry from proving that LLMs can perform trustworthy science. Notably, these prototypal automated workflows struggle with specialized scientific reasoning, long-term iterative planning, and critical analysis in collaborative workflows (12). Fully integrating AI systems in scientific discovery workflows can result in critical epistemic risks for scientific research. Unlike in test scenarios, real data is often incomplete or imprecise, which can result in emergent errors in data analysis and conclusion drawing. Giving AI systems access to physical experimental setups can lead to hazardous consequences resulting from departures from intended safety measures. Generally speaking, outsourcing more agency to these systems allows for the possibility of biasing research output, propagating unreliable results, and experimental system failures (13). To ensure that the automation of high-level human-interpretable scientific discovery tasks by AI scientists will not lead to emergent risks to scientific integrity, scientists must be able to clearly define each task in a measurable manner that is parseable and executable by an LLM.

The challenge arises from the fact that scientific discovery at a high level is not easily discretized. More explicitly, high-level human-interpretable tasks within scientific discovery workflows such as literature review or experimental design are poorly mapped to the space of low-level LLM-executable tasks such as searching or fact checking. Thus, any misalignment between the AI scientists' output and the desired outcome stems from the former's inability to fully conceptualize and comprehend the intent behind any step in the scientific discovery cycle, fundamentally a construct validity issue (14). As a rational call for better guardrailing, we posit that the danger of AI scientists comes not from simply using them, but using them without sufficient comprehension of our expectations for a task. No tool is ever optimal, but by measuring its deviation from the ideal, we can bound our expectations through quantifiable uncertainties.

Benchmarks serve as the primary method by which researchers can study the performance and limitations of AI systems on complex tasks. These evaluation workflows typically consist of a cultivated dataset and a series of metrics to test the aptitude of AI systems on a taxonomized list of tasks (14). As the concept of collaborating with AI scientists transitions from prototypal to commonplace, it becomes increasingly vital to design realistic benchmarks to measure the capacity that AI scientists have to perform common tasks in the scientific discovery workflow. However, scientific discovery cannot be reduced to a singular LLM-executable task, but is rather a dynamic workflow of indeterminate human-interpretable steps tied together via highly complex reasoning and perception that constantly evolves as new technologies are introduced. Many existing efforts to benchmark AI scientists side-step this difference by evaluating human-interpretable tasks via constructing and measuring performance on limited and potentially misleading reductions that are unrealistic and overall unrepresentative of the original task. For example, in attempting to assess citation accuracy, the CiteME benchmark is comprised of queries to identify a referenced paper given a text excerpt with a masked in-text citation (15). While an improvement over earlier retrieval-based benchmarks (16), this task construction is not representative of realistic use cases of AI systems being used as research assistants and thus indicative of a construct validity issue. It should be noted that concerns over construct validity are not exclusive to AI scientist benchmarks. Other domains involving complex tasks such as legal reasoning face similar misalignment obstacles between claims of general task mastery and the representativeness of the benchmark's task set (17).

To address the construct validity issue for AI scientist benchmarks, we propose a framework for developing benchmark tasks representative of aspects of scientific discovery derived from Arthur Koestler's descriptions of holons in *The Ghost in the Machine* (18). Our framework seeks to evaluate performance on human-interpretable tasks by summing over assessments of LLM-executable sub-tasks or holons. Each holon thus defines its own problem space to be evaluated, but also a part of a larger system to measure more general reasoning capabilities. By constructing holonic benchmarks that are representative of specific intents for scientific research tasks, yet are complete evaluations in their own rights, we can strive for collaborative settings in which AI scientists will interpret any high-level task in alignment with our expectations for scientific research. We then exemplify the power of our framework by using it to sketch a benchmark for attributional accuracy and show how it

can be used to evaluate the AI scientist's upstream ability to generate realistic and useful literature reviews.

2 AI Scientists

As this paper seeks to define a new framework for benchmarking AI scientists, it is imperative that we explicitly define what types of hybrid human-AI systems we will be testing. Conventionally, the term AI scientists refers to workflows involving AI agents that autonomously perform scientific research tasks. However, as the concept is nascent, this description remains a convention rather than any strict denotation. AI systems have been used as in silico collaborators to perform a wide range of tasks from code review to hypothesis generation (12)(19). Despite their differences in scope, each setup has the potential to obfuscate the complete measurability of the scientific discovery workflow, resulting in an emergent harm to scientific integrity. Thus, to cover all possible collaborative setups, we opt for the broadest possible definition of an AI scientist, while also accounting for the spectrum of possible human steering present:

Definition 2.1. An AI scientist refers to any hybrid human-AI workflow that relies on AI beyond simple quantitative analytic purposes. We further subdivide the category into three mutually exclusive types of AI scientists based on the measure of human steering present.

1. Human-AI assistant: The research workflow remains in its traditional, human-driven form augmented with aid from AI assistants such as chatbots.
2. Human-AI agent: Specific steps of the research process are outsourced to singular agents with a human overseer. When multiple, decoupled agents are used, the burden of aggregating the results and adjudicating lies with the human overseer. Here, the singular agents are equivalent to what are conventionally known as AI scientists: autoregressive large language models guided by agentic frameworks to adopt reasoning akin to those exhibited by human researchers (19).
3. Multi-agentic system: (Nearly) all steps of the research process are outsourced to agents with a central agent overseer and a human adjudicator (19)(20).

Across multiple fields of study, scientists have already shown the potential benefits of the ""scientist-in-the-loop" collaborative paradigm" (19). Coscientist, a GPT-4 based system for autonomous chemical research, was able to successfully optimize the reaction of palladium-catalysed cross-couplings (9). CellAgent, another GPT-4 agent, has been shown to automate single-cell RNA sequencing data analysis more reliably than its contemporaries (10). PriM is a multi-agent system designed for automated materials discovery that yields high materials exploration rate without significant reduction in scientific rigor (11).

However, integrating AI into scientific research pipelines is a double-edged sword replete with significant epistemic risks. It has already been demonstrated that AI agents struggle with computational reproducibility, a critical method for maintaining the integrity of scientific research (21). Additionally, the fact that many AI scientists are built using privately trained large language models (LLMs) makes it difficult to determine any possible biases in their outputs and whether they are inherited from their training corpora (22) or programmed preference optimization schema (23). The problem is not limited to agentic AI, as baseline LLMs continue to struggle on benchmarks of mathematical reasoning (24). Beyond experimentally verified risks, AI researchers also predict several emergent epistemic risks such as over-reliance amongst human users (23) and self-preservation in conflict with human interests (25).

3 Benchmarking Scientific Research

While scientists can rigorously answer questions about the quantifiable validity of their results, explicitly defining the required research steps for a scientific discovery proves to be more of an art. Due to the complicated reality of scientific discovery, categorizations of scientific research tasks such as the "scientific method" are typically left as instructional tools and thus vary greatly in diction (26). However, if we as scientists cannot establish a reasonable list of scientific research tasks and their aims, expecting AI scientists to perform accurately on these high-level human-interpretable tasks is

overstepping the mark. Thus, for the sake of clarity, we define the following as the research tasks required for scientific discovery:

- Literature review: This task entails detailed and extensive knowledge retrieval and contextualization within the current status quo of the field of study.
- Problem specification: This task describes the identification of a novel knowledge gap or unexplained observation to be formulated into a research project.
- Research design: This task involves the theorizing of a potential solution to the selected research problem that is both empirically testable and potentially falsifiable.
- Experimentation: This task encompasses all aspects of experimental design and execution, whether it be computational or physical in nature.
- Communication: This task represents steps required for the dissemination of scientific results via the peer review publication process.

It should be noted however that this task enumeration is not meant to define mutually exclusive tasks or be indicative of a proper chronology for scientific discovery as in practice, these research tasks are often done in parallel with many interwoven dependencies.

Scientific research tasks pose significant challenges when serving as the target for AI benchmarks. Having been in use since the 1970s (14), AI benchmarks have acquired a conventional construction consisting of a training dataset, queries, and a series of evaluation metrics to assess an AI model on certain tasks. Here, tasks refer to any mapping between a problem and action space with intensional (human-interpretable task description) and extensional (empirical pairs of states and actions) definitions (27). However, unlike strictly computational tasks with defined problem-action mappings such as image classification (28), scientific research tasks are ill-defined in scope, frequently with problem spaces mapping to other problem spaces before any notion of actions are relevant. With unclear construction, such high-level hierarchical tasks introduce significant variability amongst the possible trajectories from original problem space to outcome states on account of the actor's interpretation of the task (27). While humans can sort through the interconnected and multifaceted dependencies of scientific research tasks intuitively, it is currently unknown to what degree AI scientists could. While highlighted here for AI scientist benchmarks, this mismatch between intensional and extensional definitions manifests as construct validity issues in any assessment of language competence (14) including but not limited to legal reasoning (17), diagnosing mental disorders (29), and financial support (30). Without clear and specific definitions of outcome expectations, AI systems are prone to misalignment with our desires as collaborators.

3.1 Holonic Solution

A critical limitation of benchmarks aimed at assessing the capability of AI systems to perform language tasks is the conceptual gap between the evaluation framework and the intended human-interpretable task to be completed (31). The scientific discovery workflow is no exception. Due to the inherent uniqueness of scientific research, any task enumerated in Section 3 is in practice a test of general scientific reasoning with numerous sub-tasks each with a series of ad hoc parameters modulating the possible outcomes. This variability poses a fundamental dilemma to the construction of realistic benchmark tasks to close the aforementioned conceptual gap. Conventional benchmarks opt for reducing human-interpretable tasks to simpler versions with easy representative datasets to compile (31). However, these reduced tasks frequently end up being contrived "samples of convenience: tasks and collections of tasks arbitrarily built out of what is easily available to the team developing these benchmarks, even if such constructions are theoretically unsound" (14). For example, instead of directly tackling the human-interpretable task of literature review, the CiteME benchmark focuses on evaluating an LLM's capability for citation attribution. Defined as the task in which "a system is asked to fetch the title of a referenced paper," CiteME chooses to model queries of citation attribution as a retrieval task for a referenced paper given a text excerpt with a masked in-text citation (15). While the intensional and extensional definitions are consistent, their construction is not representative of real attributional queries by human researchers, who usually are not validating the known citations in text excerpts. Without significant attempts to realistically reduce the semantic distance between real human-interpretable tasks and the simulated tasks in the benchmark, the conventional development framework fails to accurately specify the expected

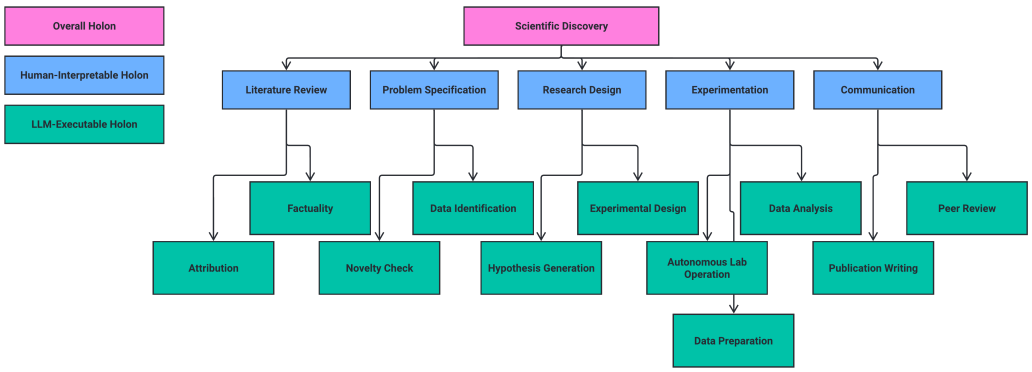

Figure 1: Prototypal holonic breakdown of the scientific discovery workflow

trajectory and outcome of the human-interpretable task and by extension, thus fails to accurately evaluate the model on its aptitude for scientific research.

To address this construct validity issue, we propose a new framework for developing benchmarks for AI scientists based on a holonic perspective on scientific research. In *The Ghost in the Machine*, Arthur Koestler defined a holon as an identifiable subsystem with a unique identity such that it is both a self-contained system in its own right as well as part of a larger holarchical structure (18). We posit that scientific research as a whole can be interpreted as a holarchical structure with the largest holon being the task of scientific discovery, further divided into intermediate holonic human-interpretable tasks like those enumerated in Section 3. These human-interpretable holons can then be subdivided into holonic tasks that are parseable and executable by LLMs. Rather than constructing a single benchmark for each human-interpretable task, our framework poses that benchmarks should be built at the LLM-executable level. By only building benchmarks for LLM-executable tasks, we minimize the conceptual gap between the evaluation framework (extensional definition) and the intensional definition of the task. Under this methodology, we rely on the combined evaluations of each LLM-executable benchmark to approximate the AI scientist's aptitude for any (human-interpretable) scientific research task. While the constituent holons themselves are still narrowly defined, by creating summative evaluation frameworks at the human-interpretable level, we become more adept at validating trajectories of AI scientists between problem spaces and subspaces for the complex tasks described in Section 3. Figure 3.1 illustrates the landscape of scientific discovery and its constituent holons when interpreted under our framework. Section 4 will exemplify the utility of our framework in practice by outlining a benchmark design to assess the attributional aptitude of AI scientists.

Koestler notes that regardless of the holarchical structure, "constituent holons are defined by fixed rules and flexible strategies" (18). As benchmarks in their own right, the LLM-executable holons are subject to the same validity issues as contemporary benchmarks. As such, we require each LLM-executable benchmark to exhibit the following criteria inspired by the forms of validity addressed in Salaudeen et al.(31):

1. Focused: The central task(s) of the benchmark should be highly specialized in scope to maintain high reproducibility and reliability.
2. Realistic: The task(s) should be constructed in accordance with actual use cases sourced from human researchers.
3. Measurable in depth: The benchmark should evaluate its task(s) with a variety of metrics, not only for technical accuracy but also for alignment and robustness against adversarial use cases.
4. Scale independence: The task(s) should be measurable across all AI scientist workflows regardless of the degree of human steering present.
5. Domain generalizability: The benchmark dataset should contain queries from multiple scientific domains with domain-specific benchmarks requiring stronger justification.

4 Attribution

To demonstrate the utility of our framework, we prototype a benchmark for a single LLM-executable holon, attribution, using it to assess an aspect of the human-interpretable task of generating realistic and useful literature reviews.

Within the system of science, literature reviewing is the mechanism by which the scientific community can referentially engage with each other by way of the compendium of scientific knowledge. While the importance of literature reviews is commonly overshadowed by experimentation or evaluation, it remains a vital foundation for scientific discovery. For researchers, the aim of literature reviews is to survey relevant literature to engender original ideas and logically support one's progression through the scientific discovery workflow. The underlying principle of measurability extends not only to empirical observations of the physical world, but also the congruence of one's work with the existing literature. Consequently, as the desire to fully automate the scientific discovery workflow escalates, we must advocate for the construction of realistic benchmarks for all constituent tasks including literature review. For without ensuring accurate literature reviews, how can we hope to protect the integrity of science?

Literature reviews seek to survey and analyze scientific literature in order to "synthesise current knowledge, critically discuss existing proposals, and identify trends" (3). As such, accurate literature reviews should be well-defined in scope, justified in its inclusion and exclusion criteria, comprehensive, and unbiased (32). Each of these criteria further complicate the already challenging task for human researchers, let alone for AI scientists. We argue that the problem space defined by the act of generating accurate and useful literature reviews proves to be too large to easily map onto action spaces in one step. This inevitably leads to a conceptual gap between the intended human-interpretable task of literature review and any attempt at designing a representative benchmark, therefore denoting a construct validity issue.

To exemplify the validity issues present with the conventional approach to benchmarking AI scientists, we review contemporary benchmarks that seek to evaluate similarly defined tasks. CiteME is a citation-based benchmark that focuses on asking whether LLMs are capable of correctly identifying a target paper from a text excerpt from a source paper including a reference to the target (15). However, the citation attribution task designed for CiteME is a limited construction. The task is an ex post facto verification of attribution, essentially side-stepping the greater issue for a question of semantic searching. As such, we find the task definition in the CiteME benchmark as not realistic. LitSearch is a retrieval benchmark that primarily seeks to assess non-LLM retrieval systems on citation recommendation queries (33). The task construction of LitSearch differs from CiteME by sampling inline target citations from source papers and then using GPT-4 to generate natural language information requests to guide its citation recommendations. LitSearch also manually reviewed the queries to ensure they were rigorous by removing any that were too close semantically to the target paper title. However, LitSearch was aimed at non-LLM retrieval systems, reducing its applicability in evaluating a broad spectrum of AI scientists. Additionally, the datasets of both CiteME and LitSearch only contain queries pertaining to ML and NLP papers specifically.

While CiteME and LitSearch struggled to overcome the validity issues posed by benchmarking the human-interpretable task of literature review, it is through such complex task setups that our proposed holonic framework for benchmark development shines. Using the conventions laid out in Section 3.1, instead of directly constructing a benchmark for literature review, we divide the human-interpretable holon of literature review into the LLM-executable holons of factuality and attribution. Figure 4 illustrates the application of our proposed framework to the human-interpretable task of literature review. Factuality refers to queries of content accuracy. A factuality benchmark would focus on assessing that summaries of scientific works present content that is accurate with respect to the modern knowledge ecosystem of a scientific discipline. Attribution refers to queries of content mapping aptitude. An attributional benchmark would therefore seek to evaluate both the relevance of provided source materials as well as the accuracy of the mapping between intellectual content to its original source.

Accurate attribution is a requirement for properly situating one's work and potential future ideas within the existing knowledge ecosystem of scientific knowledge. Beyond the primary researcher, attribution provides the only public avenue for tracing the provenance of any scientific result via the listed citations in a peer-reviewed publication. If these citations are unethically compiled, peer

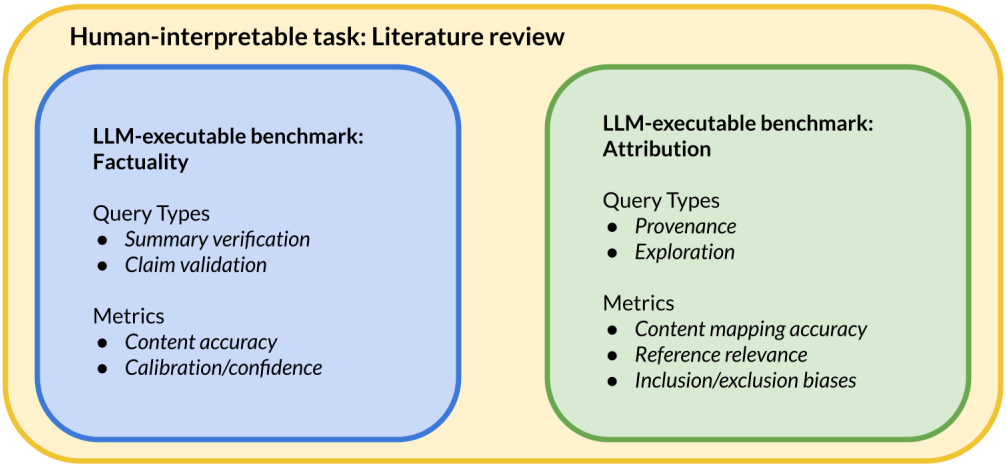

Figure 2: Proposed holonic structure for benchmarking literature review

scientists lose a critical means by which to assess the results of another party as well as a starting point for their own literature reviews. Attribution however is not limited to assessments of content mapping accuracy, but can extend to more qualitative requirements. For example, in performing a literature review, scientists need to acquire information from sources that are modern, unbiased, and diverse in methodology (32). Thus, an AI scientist would also need to recommend a variety of sources that allow for a comprehensive survey of a topic landscape. As such, we extend the definition of attributional accuracy to encompass both evaluations of accuracy and unbiased utility.

4.1 Benchmark Design

Having outlined our holonic framework for developing benchmarks for AI scientists, we now exemplify its potential by sketching out a design for a benchmark at the LLM-executable level for attributional accuracy. While prototyped on the human-interpretable holon of literature review, our framework is generalizable to any attempt at benchmarking complex, hierarchical language tasks, including but not limited to the other tasks of the scientific discovery workflow.

Having broken down the human-interpretable task of literature review to the LLM-executable level, the tasks at the holonic level prove to be specialized enough to ensure that each benchmark is focused. For the attributional benchmark, each query will be constructed as a request for a set of sources relevant to provided information in the prompt. The dataset will then consist of pairings of prompts defining the request space and appropriate listings of relevant sources defining the output space. In order for our queries to be realistic, we will request feedback from human scientists who either currently collaborate with or are interested in collaborating with AI scientists. This feedback will guide our construction of the queries that will serve as the underlying dataset by aligning the diction and syntax of these queries with real or desired use cases. Additionally, we will require that the queries sample sources from a variety of scientific domains.

We identify two distinct types of queries based on the size of the desired output space: provenance and exploration. The task of provenance is defined as a mapping between a request space populated by information requests and an output space of low dimensionality (O($k \approx 1$)) populated by potential references. The query will be framed as either directly asking for a specific source or requesting specific information that maps to a low number of potential sources. The output of provenance tasks will be structured as a ranked list of k sources, ordered based on their relevance to their initial query. Similarly to provenance queries, the task of exploration is defined as a mapping between a request space populated by information requests and an output space of unbounded dimensionality populated by relevant references. The final iteration of the task will be also structured as a ranked list of k sources, where a reasonable k will be chosen to represent typical reference counts requested by human researchers.

Despite the focused task set, we can maximize the utility of the benchmark by assessing each task with a variety of metrics. Specifically, we will choose metrics to determine the likelihood and severity

Table 1: Taxonomized list of attributional malpractices

Risk Areas		**Malpractices**
Mistake		Fabrication of content (plagiarism) Fabrication of sources Mis-mapping between sources
Misuse		Injected biases Direct output sculpting
Misalignment	Misinterpretation	Source incomprehension Source misrepresentation
	Biased Attribution	Intrinsic biases Learned strategies

by which an AI scientist could exhibit attributional malpractices. We define three risk areas for attributional malpractices: mistakes (technical failures), misuse (adversarial exploitation by human collaborators), and misalignment (knowingly acting against the intent of human collaborators). Table 1 presents a taxonomy of attributional malpractices segregated between the three risk areas. These malpractices could be measured by assessing the impact of propagated biases, model robustness to prompt perturbations, and through adversarial simulations (34). We will also test our benchmark across the multiple scales of AI scientist workflows defined in Section 2 to quantify their potential risk factors.

5 Conclusions

AI scientists have the potential to completely revolutionize the pursuit of scientific truths forever. With explosive rates of publications, the rise of the big data paradigm, and a growing desire for automation in general, it is no surprise that scientists across multiple fields of study have sought to harness the power of LLMs to augment their capabilities for scientific discovery. As such, it is a futile effort to obstruct the oncoming wave of AI scientist efforts. Clearly, we have already seen preliminary benefits from human-AI collaborations (9)(10)(11). However, the immense power of AI scientists should beget an even greater responsibility amongst current scientists to protect the integrity of scientific research.

In order to determine the potential risks posed by AI scientists to the scientific discovery workflow, we must act scientifically by measuring the aptitude of AI scientists on scientific research tasks. Unfortunately, scientific research is an ever-evolving series of ill-defined tasks that prove challenging even for human researchers to navigate. The conceptual gap posed by improper communication of the expectations for a task's outcome to an AI scientist can lead to critical failures in workflows it was designed to accelerate. In attempting to benchmark tasks at this level, this results in construct validity issues (14). Contemporary benchmarks fail to overcome this challenge, opting to generate task sets reduced in scope that are more manageable and thus more easily measurable. However, these reduced tasks frequently are not representative of their intended model, leading to construct validity issues. Our paper proposes a new framework for developing benchmarks based on Arthur Koestler's holons (18). Instead of building benchmarks for human-interpretable tasks such as literature review, we opt to construct benchmarks for LLM-parseable holons, whose semantic sum is equivalent to the original human-interpretable task. Each LLM-parseable holon is robust enough to warrant its own benchmark, but specific enough to ensure the reproducibility and reliability of the results.

To exemplify the power of our framework, we sketched a prototypal benchmark for attribution, an LLM-executable holon of the human-interpretable task of literature review. When completed, the benchmark promises to be the first benchmark to assess the attributional accuracy of AI scientists at all scales of human steering with cross-disciplinary queries. In conjunction with this paper, we hope that the benchmark will inspire further efforts to benchmark AI scientists as they become ubiquitous additions to the scientific discovery workflow.

References

[1] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

[2] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **79**, 2554–2558 (1982). URL https://www.pnas.org/doi/10.1073/pnas.79.8.2554. Publisher: Proceedings of the National Academy of Sciences.

[3] de la Torre-López, J., Ramírez, A. & Romero, J. R. Artificial intelligence to automate the systematic review of scientific literature. *Computing* **105**, 2171–2194 (2023). URL https://doi.org/10.1007/s00607-023-01181-x.

[4] Wang, Q., Downey, D., Ji, H. & Hope, T. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 279–299 (2024). URL http://arxiv.org/abs/2305.14259. ArXiv:2305.14259 [cs].

[5] Si, C., Yang, D. & Hashimoto, T. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers (2024). URL http://arxiv.org/abs/2409.04109. ArXiv:2409.04109 [cs].

[6] Yamada, Y. *et al.* The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search (2025). URL http://arxiv.org/abs/2504.08066. ArXiv:2504.08066 [cs].

[7] Alkan, A. K. *et al.* A Survey on Hypothesis Generation for Scientific Discovery in the Era of Large Language Models (2025). URL http://arxiv.org/abs/2504.05496. ArXiv:2504.05496 [cs] version: 1.

[8] Arlt, S. *et al.* Meta-Designing Quantum Experiments with Language Models (2025). URL http://arxiv.org/abs/2406.02470. ArXiv:2406.02470 [quant-ph].

[9] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023). URL https://www.nature.com/articles/s41586-023-06792-0. Publisher: Nature Publishing Group.

[10] Xiao, Y. *et al.* CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis (2024). URL http://arxiv.org/abs/2407.09811. ArXiv:2407.09811 [cs].

[11] Lai, Z. & Pu, Y. PriM: Principle-Inspired Material Discovery through Multi-Agent Collaboration (2025). URL http://arxiv.org/abs/2504.08810. ArXiv:2504.08810 [cs] version: 1.

[12] Reddy, C. K. & Shojaee, P. Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges (2024). URL http://arxiv.org/abs/2412.11427. ArXiv:2412.11427 [cs] version: 2.

[13] Gridach, M., Nanavati, J., Abidine, K. Z. E., Mendes, L. & Mack, C. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions (2025). URL http://arxiv.org/abs/2503.08979. ArXiv:2503.08979 [cs].

[14] Raji, I. D., Bender, E. M., Paullada, A., Denton, E. & Hanna, A. AI and the Everything in the Whole Wide World Benchmark (2021). URL http://arxiv.org/abs/2111.15366. ArXiv:2111.15366 [cs].

[15] Press, O. *et al.* CiteME: Can Language Models Accurately Cite Scientific Claims? (2024). URL http://arxiv.org/abs/2407.12861. ArXiv:2407.12861 [cs].

[16] Gu, N., Gao, Y. & Hahnloser, R. H. R. Local Citation Recommendation with Hierarchical-Attention Text Encoder and SciBERT-based Reranking (2022). URL http://arxiv.org/abs/2112.01206. ArXiv:2112.01206 [cs].

[17] Guha, N. *et al.* LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models (2023). URL http://arxiv.org/abs/2308.11462. ArXiv:2308.11462 [cs].

[18] Koestler, A. *The Ghost in the Machine*. An Arkana book : philosophy (Penguin Group (USA) Incorporated, 1989). URL https://books.google.com/books?id=tJS-QgAACAAJ.

[19] Gottweis, J. *et al.* Towards an AI co-scientist (2025). URL http://arxiv.org/abs/2502.18864. ArXiv:2502.18864 [cs].

[20] Yager, K. G. Towards a science exocortex. *Digital Discovery* **3**, 1933–1957 (2024). URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00178h. Publisher: RSC.

[21] Siegel, Z. S., Kapoor, S., Nagdir, N., Stroebl, B. & Narayanan, A. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark (2024). URL http://arxiv.org/abs/2409.11363. ArXiv:2409.11363 [cs].

[22] Wal, O. v. d. *et al.* Undesirable Biases in NLP: Addressing Challenges of Measurement. *Journal of Artificial Intelligence Research* **79**, 1–40 (2024). URL http://arxiv.org/abs/2211.13709. ArXiv:2211.13709 [cs].

[23] Mitchell, M., Ghosh, A., Luccioni, A. S. & Pistilli, G. Fully Autonomous AI Agents Should Not be Developed (2025). URL http://arxiv.org/abs/2502.02649. ArXiv:2502.02649 [cs] version: 2.

[24] Satpute, A. *et al.* Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange (2024). URL http://arxiv.org/abs/2404.00344. ArXiv:2404.00344 [cs] version: 1.

[25] Bengio, Y. *et al.* Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? (2025). URL http://arxiv.org/abs/2502.15657. ArXiv:2502.15657 [cs].

[26] Fontes, D. T. M. Nature of Science in the classroom: empirical insights into investigating physics. *A Física na Escola* **22**, 230136 (2024). URL http://arxiv.org/abs/2504.03912. ArXiv:2504.03912 [physics].

[27] Schlangen, D. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In Zong, C., Xia, F., Li, W. & Navigli, R. (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 670–674 (Association for Computational Linguistics, Online, 2021). URL https://aclanthology.org/2021.acl-short.85/.

[28] Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge (2015). URL http://arxiv.org/abs/1409.0575. ArXiv:1409.0575 [cs].

[29] Sun, J. *et al.* Practical AI application in psychiatry: historical review and future directions. *Molecular Psychiatry* **30**, 4399–4408 (2025). URL https://www.nature.com/articles/s41380-025-03072-3. Publisher: Nature Publishing Group.

[30] Vuković, D. B., Dekpo-Adza, S. & Matović, S. AI integration in financial services: a systematic review of trends and regulatory challenges. *Humanities and Social Sciences Communications* **12**, 562 (2025). URL https://www.nature.com/articles/s41599-025-04850-8. Publisher: Palgrave.

[31] Salaudeen, O. *et al.* Measurement to Meaning: A Validity-Centered Framework for AI Evaluation (2025). URL http://arxiv.org/abs/2505.10573. ArXiv:2505.10573 [cs] version: 4.

[32] Ressing, M., Blettner, M. & Klug, S. J. Systematic Literature Reviews and Meta-Analyses. *Deutsches Ärzteblatt International* **106**, 456–463 (2009). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2719096/.

[33] Ajith, A. *et al.* LitSearch: A Retrieval Benchmark for Scientific Literature Search (2024). URL http://arxiv.org/abs/2407.18940. ArXiv:2407.18940 [cs].

[34] Lin, L. *et al.* Against The Achilles' Heel: A Survey on Red Teaming for Generative Models (2024). URL http://arxiv.org/abs/2404.00629. ArXiv:2404.00629 [cs] version: 2.